Test-retest reliability of a new delay aversion task and executive function measure.

Kunsti, J.; Stevenson, J.; Oosterlaan, J.; Sonuga-Barke, J.S.

published in
British Journal of Developmental Psychology
2001

DOI (link to publisher)
10.1348/026151001166137

document version
Publisher's PDF, also known as Version of record

British Journal of Developmental Psychology (2001), 19, 339–348 *Printed in Great Britain*

Test–retest reliability of a new delay aversion task and executive function measures

Jonna Kuntsi*
Behavioural Sciences Unit, Institute of Child Health, University College London Medical School, UK

Jim Stevenson
Centre for Research into Psychological Development, University of Southampton, UK

Jaap Oosterlaan
Department of Clinical Neuropsychology, Free University, Amsterdam, The Netherlands

Edmund J. S. Sonuga-Barke
Centre for Research into Psychological Development, University of Southampton, UK

Despite the wide adoption of measures of executive functions and motivational tendencies in studies of developmental disorders and child psychopathology, few studies have investigated their test–retest reliability. The present paper examines the reliability of a new measure of delay aversion, three measures of working memory, a response inhibition measure and a measure of dual task performance. The children, aged between 7 and 15 years, performed the tasks twice, with a 2-week period in between the sessions. Using a relatively conservative criterion, only the delay aversion task and one of the working memory measures (delayed response alternation) demonstrated satisfactory test–retest reliability. The other two working memory measures (sentence span and counting span) showed modest reliability. For the inhibition measure (stop task) the results were mixed, with poor to modest reliabilities obtained for the various derived measures. The dual task failed to demonstrate adequate test–retest reliability. These differential reliabilities need to be borne in mind when interpreting the results of studies using these measures. In particular the effect of low reliability on statistical power and the Type II error rate should be considered.

Measures of 'executive functions', such as response inhibition and working memory, and of motivational tendencies are often used in studies of developmental disorders and child psychopathology. However, few studies have investigated their test–retest reliability.

* Requests for reprints should be addressed to Jonna Kuntsi, Behavioural Sciences Unit, Institute of Child Health, University College London Medical School, 30 Guilford Street, London WC1N 1EH, UK (e-mail: j.kuntsi@ich.ucl.ac.uk).

Response inhibition refers in particular to the ability to withhold a prepotent response (Barkley, 1997). A laboratory analogue, called the stop signal paradigm, provides an empirical measure of the ability to interrupt an ongoing response (Logan & Cowan, 1984; Logan, Cowan, & Davis, 1984). This paradigm has several advantages over other measures of inhibition. First, the stop task is based on an explicit model of the inhibitory process. Second, it allows a distinction between inhibitory control and the processes involved in the execution of the primary task (a two-choice reaction time task). Although successful inhibition of the ongoing action is not observable, the stop task provides a way to measure stop signal reaction time.

Kindlon, Mezzacappa, and Earls (1995) investigated the temporal stability of the stop task. The children (N = 31; ages 6–16) who participated in the study were recruited from schools for children with externalizing behaviour disorders. The period between the test and retest sessions varied between 2 and 5 months. The results showed moderate to high stability (bivariate correlations .61–.79; squared partial correlations, controlling for age, .33–.52) for all the stop task variables which were included in the study. Different studies have used markedly different versions of the stop task (Oosterlaan, Logan, & Sergeant, 1998), however, and we are not aware of reliability data having been published for any other versions of the task.

Working memory can be defined as 'a limited capacity computational arena' (Pennington, 1994, p. 248). Working memory maintains representations of past, present and future briefly over time, 'in a common system so they can interact' (p. 248). It is future oriented and transient. A variety of working memory measures have been developed. One of the latest working memory measures is the delayed response alternation task (DRA), developed by Gold and colleagues (Gold, Faith Berman, Randolph, Goldberg, & Weinberger, 1996). Working memory performance can also be assessed with pencil-and-paper tasks such as the sentence span (Daneman & Carpenter, 1980) and counting span (Case, Kurland, & Goldberg, 1982) tasks. To our knowledge no reliability data have been published previously for these tasks.

Another measurement approach to executive functions which emphasizes the role of working memory is the dual task approach. However, Baddeley and colleagues (Baddeley, Della Sala, Gray, Papagno, & Spinnler, 1997) point out that 'dual task performance is intended to be an indicator not of working memory nor of motor dexterity per se but of the patient's ability to deploy the available resources of memory capacity' (p. 72). Baddeley *et al.* (1997) recently developed a new test of dual task performance. In their preliminary investigations, Baddeley *et al.* found the task to be a very promising measure of central executive functioning, but they raised some concern over whether the task demonstrates adequate reliability. A Pearson correlation of only .44 was obtained between first and second occasions of testing based on scores from 33 adults. No data were collected on children's performance on the task.

A different approach to assessing children's functioning is the tasks which focus on motivational tendencies. Within the hyperactivity literature the delay aversion theory (e.g. Sonuga-Barke, 1994; Sonuga-Barke, Taylor, Sembi, & Smith, 1992) exemplifies such an approach. This theory argues that hyperactive children are not impulsive in the sense that they would be *unable* to wait, although they usually *prefer* not to wait. On a task in which the children had to make a choice between a small immediate reward and a large delayed reward, hyperactive children chose the small immediate reward more

often than control children, earning fewer points, *only* when this strategy reduced the overall delay period (Sonuga-Barke *et al.*, 1992). When choosing the small immediate reward lead to a post-reward delay, such that the overall delay was the same as when choosing the large delayed reward, the hyperactive children waited as well as the control children for the larger reward. Previous studies have not established the test–retest reliability of the delay aversion tasks. In this paper we introduce a new task called the Maudsley Index of Childhood Delay Aversion.

The study reported here aimed to establish the test–retest reliability of a battery of tasks measuring inhibition, working memory, dual task performance and delay aversion.

Method

Samples and procedure

Sample 1. The test–retest reliability for the dual task, the counting span and sentence span tasks was carried out in two different inner London schools, a primary and a secondary school. We asked the headteachers in both schools to write a letter to the parents of a representative sample of children, in terms of age and gender, asking for permission for their child to take part. The parents of only one child of those contacted refused to allow their child to participate. In addition, one child, while given parental consent, did not wish to take part in the study.

A total of 34 children, 15 girls and 19 boys, participated. The children ranged in age from 7.9 to 15.3 years (mean age = 11.4 years, SD = 2.3 years). Twenty of the children were from the primary school and 14 from the secondary school. The majority (71%) of them were Caucasian, 15% were Indian or Pakistani, 3% were Asian, 9% were African/Caribbean and 3% were classified as 'other' in terms of ethnic origin.

The children were tested individually in a separate room in the school. On any single day, two testers (out of three) assessed the children simultaneously. There were two fixed orders of task administration, both administered with equal frequency. The children were tested again after a 2-week period. The tests were presented in the same order for each child as they had been presented at time 1. However, each tester now assessed the children the other tester had assessed previously.

Sample 2. A different sample participated in the test–retest reliability of three tasks presented on a computer: the stop task, the delayed response alternation (DRA) task and the delay aversion task. We wrote instructions for the new delay aversion task and made some revisions to the instructions for the DRA and stop tasks. The instructions for the DRA were written for adults and were therefore inappropriate for children. A professional translator translated the stop task instructions from Dutch into English.

The same three testers administered the tasks in an inner London primary school. The headteacher in the school wrote to the parents of children in the 7–11 age range, asking for permission for their child to take part. We then chose the children to be tested from those whose parents had given their consent so as to obtain approximately equal numbers of girls and boys and children of different ages. Within each of these 'subgroups' we chose the children randomly. The sample consisted of 18 children: 8 girls and 10 boys. Mean age of the children was 8.8 years (SD = 1.4 years). In terms of ethnic origin, the majority (78%) of the children were Caucasian, 11% were Asian and 11% were African/Caribbean.

On any single day, one tester assessed children individually in a separate room in the school. The order of task administration was not fixed across children, but the tasks were administered in the same order at test and retest for each child. There was a 2-week period in between the test and retest sessions. A different examiner tested each child at time 1 and time 2.

Measures

Sentence span (Daneman & Carpenter, 1980 – the original version; Siegel & Ryan, 1989 – the version used in this study) and counting span (Case et al., 1982) tasks. These tasks are working memory measures. In the

sentence span task, the tester read sentences out to the child who had to supply the missing last word for each sentence. At the end of each set, the child was asked to repeat all the words that he or she had supplied, in the correct order. The tester first gave the child a practice sentence and then, in order also to practise recalling the supplied words, two further sentences. The task proper began with two-sentence sets and, unless the child failed all three sets of any level, finished with five-sentence sets. The sentences for the task were chosen so that the missing word was virtually predetermined. We made some modifications to the sentences in order for them to be more appropriate for British children. For example, the sentence 'In a baseball game the pitcher throws the –' was changed into 'In a tennis game the player hits the —'.

The counting span task is similar to the sentence span task except that the child was asked to count dots on cards rather than to supply words. Each card consisted of a random arrangement of yellow (targets) and blue dots. The blue dots were arranged randomly among the yellow dots to prevent counting by subitizing. The size of the cards was 14cm × 21cm and the dots were 0.9cm in diameter. There were four levels of difficulty, from two-card sets to five-card sets. At each level of difficulty, there were three sets of cards. The tester asked the child to touch each yellow dot with his or her finger and to count out loud. The practice started with counting the yellow dots on one card. The tester then, presenting one card at a time, asked the child to count the dots on two cards and, when presenting a blank card, to recall the numbers of dots on the previous cards. The testing proper started with two-card sets and, unless the child failed all three sets of any level, finished with five-card sets. The possible scores range from 0 to 12 on both the sentence and counting span tasks, with one point being earned for each correct set.

Delayed response alternation task (DRA: Carpenter & Gold, 1994; Gold et al., 1996). This task is a computerized working memory measure, in which the child tried to find out a rule that determined which of two stimuli was the correct choice. Two boxes, one coloured (yellow) and the other uncoloured (with only the outline drawn), were first presented on the screen for 1 second. After a 2-second presentation of an empty screen, two uncoloured boxes appeared on the screen and the child had to choose one of these boxes, either the one on the side where the yellow box had been or the one on the side where the uncoloured box had been. The computer gave feedback as to whether the choice was correct or incorrect (the word right or wrong was presented on the screen for 1.5 seconds immediately after the child had responded). New stimuli (another two boxes, one yellow and the other uncoloured) then appeared on the screen after a 1.5 seconds delay.

The task for the child was to find out the rule that the computer used to decide which box was the correct one each time. If the child did not find out the rule on his or her own, the rule was then taught explicitly. The rule involved choosing the yellow and the uncoloured box on alternate trials, regardless of position. The position of the yellow box varied randomly. All children performed the task twice, regardless of whether they found out the rule on their own.

Before the children started the task proper, they first practised responding (pressing the numbers 1 and 2 on the keyboard) with a practice version of the task. In this practice version the correct rule was always to choose the coloured (blue) box. The children were told after the practice that the rule might be different in the 'real game'.

Two variables are obtained from the DRA task: the percentage of correct choices 'pre and post instruction'. However, as many children would be expected to remember the rule at restest, the comparison of interest in terms of test–retest reliability was that between time 1 and time 2 *post instruction* scores.

Dual task. The dual task (Baddeley *et al.*, 1997) is a pencil-and-paper measure in which the participant first performs two simple tasks (a memory span task and a tracking task) separately and then simultaneously (the dual task condition).

The child's digit span was first tested using a standard procedure. The examiner read aloud lists of digits at the rate of 1 digit per second and the child was asked to repeat them in their order of presentation. The first level consisted of three 2-digit lists; the lists at each successive level were 1 item longer than the lists at the previous level. The child's digit span was taken to be the maximum length at which all three lists at a particular level were reproduced without error. The examiner then presented

continuously, over a 2-minute period, lists of digits at span length which the child had to repeat in order of presentation. The digit span score used in the calculations to obtain the dual task score (μ; see below) was the proportion of these lists correctly recalled.

The tracking task involved the child crossing out boxes (of size 1cm square) which had been linked to form a path laid out on an A4-size sheet of white paper. The examiner first gave the child several practice trials with a 10-box path. In the tracking task proper, the child was asked to follow paths through sheets with 80 boxes over a period of 2 minutes. The performance measure was the number of boxes successfully marked.

In the dual task condition the child performed the memory span task and the tracking task simultaneously. The measure of interest that one obtains from the dual task is that indexed as 'μ' (see Baddeley *et al.*, 1997, for how the measure is calculated). This measure expresses the child's dual task performance as a percentage of single task performance, the contributions from the two tasks being equally weighted.

Maudsley Index of Childhood Delay Aversion. This is a new computer task designed to test the delay aversion theory of hyperactivity. The full task involves several conditions, but in the present study we included only the condition that is predicted to distinguish children who are hyperactive from those who are not. In this task the child had to make a choice, on 20 occasions, between a small immediate reward (1 point involving a 2-second pre-reward delay) and a large delayed reward (2 points involving a 30-second pre-reward delay). If the child chose the small reward, the next trial started immediately afterwards; this of course reduced the overall length of the session.

The task was presented as a space game, in which the child, as a captain of a spaceship, had to destroy enemy spacecraft (using the computer mouse). The aim of the game was to earn as many points as possible and to motivate the children they were told that they would receive a small prize in the end (in this study the children received pencils). Before the experimental trials, the child first practised using the mouse and choosing each of the rewards. The tester also asked the child questions about the game, to ensure that he or she had understood the rules and aims of the game correctly. The delay aversion variable used in the analyses was the percentage of choices for the 2 points delayed reward.

Stop task (Logan & Cowan, 1984; Logan et al., 1984 – the original version; Oosterlaan & Sergeant, 1998 – the version used in this study). This computer task measures inhibition and is based on Logan and Cowan's (1984) 'race model' of inhibition. Each trial began with a 350 milliseconds presentation of a fixation point ('+'-sign presented at the centre of the screen). The presentation of the stimulus (an airplane, displayed for 1500 milliseconds) then followed. The inter-trial interval was 1000 milliseconds. A Keithley PIO-12 digital interface board enabled the stimuli to be presented and the data to be collected with millisecond accuracy. The stimuli appeared equally often on either side of the screen within each block and the stop signals were presented equally often after left- and right-sided presentations of the stimuli. A go trial always followed a stop trial, except once in each block where two stop signals were presented in succession.

The percentage of stop trials was 25%. The stop signals were presented equally often at each of the four stop signal intervals (50, 200, 350 and 500 milliseconds before the child's expected response). The expected moment of responding was estimated from the child's mean reaction time (MRT) in the preceding block of trials. MRT was calculated across correctly executed responses on go trials. The stop signals were 1kHz tones produced by a function generator.

The task was presented as a game in which the child had to perform tasks similar to those of an air-traffic controller. The child was first taught to respond to airplanes appearing on the computer screen by pressing the response button that was on the same side as the plane (a two-choice reaction time task). The child was then instructed to withhold responding whenever he or she heard a tone on headphones (the 'stop' trials), but otherwise to keep on responding to the planes as quickly as possible (the 'go' trials). The tones were presented at four different intervals after the presentation of the planes. All children did two practice and four experimental blocks (with 64 trials in each) on this task and were given short breaks between the blocks.

The following stop task variables were used in the analyses: mean probability of inhibition, slope of the inhibition function, stop signal reaction time (SSRT), mean reaction time (MRT), standard deviation

of reaction times (SD of RTs), total number of errors, number of omission errors and number of commission errors (see Oosterlaan *et al.*, 1998, for further details.)

Wechsler Intelligence Scales for Children (WISC-III^{UK}: Wechsler, 1992). Four subtests from the WISC were used to obtain an estimate of the child's full-scale IQ: Picture Completion, Block Design, Vocabulary and Similarities. We chose these subtests because they have high loadings on the performance and verbal IQ factors, respectively.

Results

As the measures of test–retest reliability, we calculated inter-class and intra-class Pearson product moment correlations, as well as partial correlations controlling for the effects of age, between scores on each of the measures at test and retest. Intra-class correlations are obtained from double-entered data, therefore taking into account learning effects from time 1 to time 2 (McGraw & Wong, 1996). In addition, we report *t* test results for the comparisons between mean scores at test and retest for each measure.

An acceptable level of reliability depends on the type of test used. Rust and Golombok (1989) point out that although reliabilities in excess of .9 are obtained for individual IQ tests, with personality tests reliabilities of .7 are acceptable. With the subtle tests of specific abilities in children used in this study, we consider reliabilities of .7 or higher as satisfactory, whereas reliabilities of .5 and .6 can be considered as modest. This is in agreement with the procedure Kindlon *et al.* (1995) adopted, as they similarly set acceptability at .7 for bivariate correlations (for age-standardized scores) and .45 for squared partial correlations. When determining acceptability of the test–retest reliability results, we focus on the partial correlations, to control for age effects, and the intra-class correlations.

Table 1 shows the test–retest reliability results. Of the working memory measures, the sentence span and counting span tasks demonstrated modest reliability and the DRA task (post instruction) satisfactory reliability. The *t* test results show that there were significant learning effects for both the sentence span and counting span tasks, but not for the DRA task. The children made an average of 56% correct choices on the DRA task at time 1 pre instruction and 77% post instruction. At time 2 they made an average of 76% correct choices pre instruction and 78% afterwards. More than half (61%) of the children did not find out the rule on their own at time 1, but 78% of them remembered the rule at time 2.

The test–retest reliability results were low for the dual task measure. To investigate this further, the results were also analysed separately for those measures on which the two-component measure of dual task performance (μ) is based. This more detailed analysis indicates that it was the memory span measure (single condition: interclass $r = -.11$, intra-class $r = -.12$, partial $r = -.12$; dual condition: inter-class $r = 0.13$, intra-class $r = .12$, partial $r = .13$) rather than the tracking measure (single condition: inter-class $r = .95$, intra-class $r = .87$, partial $r = .89$; dual condition: interclass $r = .89$, intra-class $r = .78$, partial $r = .81$) that was unreliable in the task.

The delay aversion task demonstrated satisfactory test–retest reliability. At time 1 the children chose the larger reward on 53.9% of the trials and at time 2 on 54.4% of the trials on average.

Of the stop task variables the *inter*-class correlation coefficients were satisfactory for

Table 1. Test–retest reliability results

Measure	Inter-class correlation	Intra-class correlation	Partial correlation[a]	Time 1		Time 2		*t* value	d.f.	*p* value
				M	SD	*M*	SD			
Sentence span	0.71	0.65	0.54	4.12	2.29	5.00	2.49	2.82	33	.008
Counting span	0.67	0.55	0.54	6.35	2.88	7.88	2.68	−3.92	33	.001
DRA post instruction	0.74	0.74	0.65	77.22	14.37	77.92	14.15	−0.28	17	.78
Dual task (μ)	0.35	0.33	0.34	94.44	8.49	96.29	7.61	−1.17	33	.25
Delay aversion	0.74	0.74	0.69	53.89	31.37	54.44	35.60	−0.10	17	.92
Stop task										
Mean probability of inhibition	0.72	0.52	0.60	62.48	12.18	70.26	10.63	−3.78	17	.001
MRT	0.66	0.35	0.31	488.72	95.87	404.58	85.44	4.75	17	.001
SD of RTs	0.74	0.64	0.45	115.94	41.39	96.25	38.76	2.90	17	.01
Total errors	0.49	0.41	0.32	4.67	5.84	7.50	8.11	−1.65	17	.12
Commission errors	0.45	0.22	0.36	2.17	2.66	4.94	6.11	−2.16	17	.05
Ommission errors	0.37	0.37	0.08	2.50	3.45	2.56	3.84	−0.06	17	.96
Slope of inhibition function	0.32	0.29	0.31	0.14	0.03	0.14	0.03	1.01	17	.33
SSRT	0.21	0.11	0.23	230.97	43.56	201.39	58.98	1.92	17	.07

[a] Controlling for age effects.

the mean probability of inhibition, standard deviation of reaction times and mean reaction time. However, the partial correlations and intra-class correlations were lower, particularly for mean reaction time. The significant *t* test results also show that there were learning effects from time 1 to time 2 testing. The test–retest correlation coefficients were lower for the error variables, slope of the inhibition function and stop signal reaction time.

We also examined possible tester effects by comparing mean scores on the measures across testers for time 1 data. The children were not allocated to testers in any systematic way and no consistent effects of group would be expected. The results from independent *t* tests were non-significant, with one exception. The only significant comparison between two testers emerged for the stop task variable of standard deviation of reaction times. To explore this finding further, we performed a similar comparison between the two testers for time 2 data for this same variable: the result was non-significant. These results suggest that the test–retest reliability results reported for the measures are not affected significantly by tester effects.

The average estimated full-scale IQ for sample 1 was surprisingly low at 83.74 (SD = 20.79), given that the sample was recruited from ordinary classrooms. We therefore decided to carry out additional correlational analyses, controlling for IQ. The partial correlations indicated that controlling for IQ did not have a noticeable effect on the reliability results (partial inter-class correlations, controlling for IQ: .69 for sentence span, .69 for counting span and .35 for 'μ'; partial intra-class correlations, controlling for IQ: .62 for sentence span, .54 for counting span and .32 for 'μ').

Discussion

For many widely used measures of children's cognitive and motivational functioning their reliability is often simply assumed rather than investigated. We aimed to rectify this situation by establishing the test–retest reliability of specific measures of response inhibition, working memory, dual task performance and delay aversion.

The delay aversion task, the Maudsley Index of Childhood Delay Aversion, showed satisfactory test–retest reliability. We have recently demonstrated that this new delay aversion task also has good discriminant validity, distinguishing between hyperactive and control groups (Kuntsi, Oosterlaan, & Stevenson, in press).

Using a relatively strict criterion (focusing on the intra-class and partial correlations), only one of the working memory measures, the delayed response alternation task, demonstrated satisfactory test–retest reliability. As only 39% of the children found out the rule on their own on this task, the pre instruction scores are not useful from the point of view of test–retest reliability. The sentence span and counting span tasks demonstrated modest reliability.

Kindlon *et al.* (1995) previously reported temporal stability data for the stop task. However, the particular version of the stop task employed in that study was different from the version we used. Focusing on the partial correlations (either squared or unsquared for *both* sets of results), the test–retest reliability results were somewhat lower in our study than in the Kindlon *et al.* (1995) study for mean probability of inhibition, standard deviation of reaction times, number of commission errors and the slope of inhibition function. Kindlon *et al.* (1995) did not report the results for mean (nonsignal)

reaction time, total number of errors, number of omission errors or stop signal reaction time.

Amongst the lowest test–retest reliabilities for the stop task variables in our study were those for the slope of the inhibition function and stop signal reaction time. These variables are derived variables and are both based on reaction times and the probability of inhibition. Such derived variables necessarily have lower reliability than the measures on which they are based, as the measurement errors are compounded in the derived variables. A possible reason why Kindlon *et al.* (1995) obtained somewhat better results for some of the stop task variables is the differences in the samples: whereas our sample was a general population sample, the sample in the Kindlon *et al.* study was obtained from schools for children with externalizing behaviour disorders. It is possible that in our sample there was less variability in scores, which would attenuate the test–retest reliabilities. This is an issue worth investigating in future studies.

The dual task failed to demonstrate adequate stability over time. A closer inspection of the results indicated that this was due to poor test–retest reliability of the memory span task. The other part of the dual task, the tracking task, showed very good reliability. The results from adults in Baddeley *et al.*'s (1997) preliminary investigations showed a similar pattern. Future studies should therefore attempt specifically to improve the reliability of the memory span part of the task.

A limitation of the study, in particular regarding the second sample, is the small sample size. The age range of the children in this sample, which focused on the computer measures, was also narrower, which may have led to less variability in scores.

In sum, the present study demonstrated satisfactory test–retest reliability for the delayed response alternation task and the new measure of delay aversion, and modest reliability for the sentence span and counting span tasks. For the stop task the results were mixed, with the highest partial correlation (controlling for age effects) reaching .6. The dual task failed to demonstrate adequate test–retest reliability. As low reliability affects statistical power and the Type II error rate, these variations in reliability need to be borne in mind when interpreting the results of studies using these measures. In future studies these test–retest reliability results should also be considered when selecting suitable tests.

Acknowledgements

We thank all the children and teachers who took part in these studies. We also thank Doug Barrett and Emma Canning for invaluable help during data collection, and Jody Warner-Rogers for her contribution to the development of the delay aversion measure. This research was funded by a Wellcome Prize Studentship to the first author.

References

Baddeley, A., Della Sala, S., Gray, C., Papagno, C., & Spinnler, H. (1997). Testing central executive functioning with a pencil-and-paper test. In P. Rabbitt (Ed.), *Methodology of frontal and executive functions* (pp. 61–80). Hove: Erlbaum.

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin, 121*, 65–94.

Carpenter, C. J., & Gold, J. M. (1994, February). *Prefrontal functioning in schizophrenia*. Paper presented at the International Neuropsychological Society Annual Meeting, Seattle.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33*, 386–404.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450–466.

Gold, J. M., Faith Berman, K., Randolph, C., Goldberg, T. E., & Weinberger, D. R. (1996). PET validation of a novel prefrontal task: Delayed response alternation. *Neuropsychology, 10*, 3–10.

Kindlon, D., Mezzacappa, E., & Earls, F. (1995). Psychometric properties of impulsivity measures: Temporal stability, validity and factor structure. *Journal of Child Psychology and Psychiatry, 36*, 645–661.

Kuntsi, J., Oosterlaan, J., & Stevenson, J. (in press). Psychological mechanisms in hyperactivity: I Response inhibition deficit, working memory impairment, delay aversion or something else? *Journal of Child Psychology and Psychiatry*.

Logan, G. D., & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review, 91*, 295–327.

Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the ability to inhibit responses in simple and choice reaction time tasks: A model and a method. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 276–291.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

Oosterlaan, J., Logan, G. D., & Sergeant, J. A. (1998). Response inhibition in AD/HD, CD, comorbid AD/HD+CD, anxious, and control children: A meta-analysis of studies with the stop task. *Journal of Child Psychology and Psychiatry, 39*, 411–425.

Oosterlaan, J., & Sergeant, J. A. (1998). Response inhibition and response re-engagement in ADHD, disruptive, anxious and normal children. *Behavioural Brain Research, 94*, 33–43.

Pennington, B. F. (1994). The working memory function of the prefrontal cortices: Implications for developmental and individual differences in cognition. In M. M. Haith, J. B. Benson, R. J. Roberts Jr, & B. F. Pennington (Eds.), *The development of future-oriented processes* (pp. 243–289). Chicago: University of Chicago Press.

Rust, R., & Golombok, S. (1989). *Modern psychometrics: The science of psychological assessment*. London: Routledge.

Siegel, L. S., & Ryan, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled children. *Child Development, 60*, 973–980.

Sonuga-Barke, E. J. S. (1994). Annotation: On dysfunction and function in psychological theories of childhood disorder. *Journal of Child Psychology and Psychiatry, 35*, 801–815.

Sonuga-Barke, E. J. S., Taylor, E., Sembi, S., & Smith, J. (1992). Hyperactivity and delay aversion: I The effect of delay on choice. *Journal of Child Psychology and Psychiatry, 33*, 387–398.

Wechsler, D. (1992). *Wechsler Intelligence Scale for Children – Third Edition, UK*. London: The Psychological Corporation.

Received 28 May 1999; revised version received 25 August 2000